ECFR: Entropy-Consistency Flow Rectification in Test-Time Adaptation for Medical Foundation Models

Shangkun Li[1,2,3], Yun Gu[1,2,3] (✉), and Hanxiao Zhang[1,4] (✉)

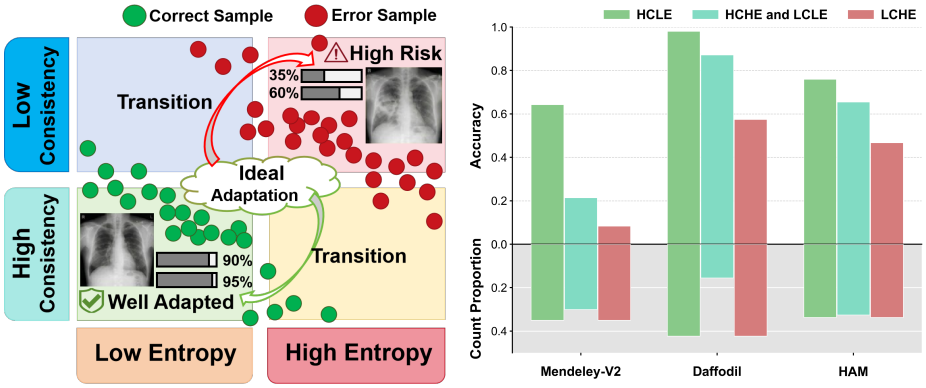
[1] Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[2] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[3] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[4] Institute of Medical Robotics & School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
yungu@ieee.org; hanxiao.zhang@sjtu.edu.cn

Abstract. Deploying medical Foundation Models (FMs) is often severely compromised by continuous clinical distribution shifts. While Test-Time Adaptation (TTA) offers a promising solution, existing paradigms rely on overly simplified, uni-directional optimization that blindly enforces confidence on flawed representations, inevitably precipitating error accumulation and model collapse. To address this critical safety gap, we propose Entropy-Consistency Flow Rectification (ECFR), which fundamentally redefines TTA as a spatial routing problem. Central to our approach are the Entropy-Consistency Quadrants, which explicitly disentangle streaming data into distinct diagnostic states. Guided by this spatial partitioning, ECFR pioneers a quadrant-wise flow rectification strategy: it consolidates potentially correct samples via minimization while actively sequestering high-risk erroneous inputs via maximization, strictly halting the propagation of erroneous gradients. Furthermore, a dynamic memory bank continuously calibrates these quadrant thresholds against distribution drift. Comprehensive evaluations on two state-of-the-art (SOTA) medical FMs across three diverse out-of-distribution (OOD) datasets demonstrate that ECFR successfully prevents model collapse, achieving superior accuracy and confidence calibration for safe clinical deployment. The code is available at https://github.com/EndoluminalSurgicalVision-IMR/ECFR.

Keywords: Medical Foundation Models · Distribution Shift · Test-Time Adaptation.

1 Introduction

Foundation Models (FMs) demonstrate remarkable potential in medical imaging, yet their clinical deployment remains challenged by real-world distribution



(a) (b)

Fig. 1. (a) Entropy-consistency analysis framework and ideal adaptation for data flow. Metric definitions and the heuristic rationale for the ideal flow direction are detailed in Sec. 2.1. (b) Comparison of accuracy across four quadrants in three OOD datasets. We partitioned the quadrants using the median values of overall dataset entropy and consistency, ensuring comparable data volumes and statistically significant accuracy (as evidenced by the count proportion).

shifts across imaging devices, institutional protocols, and patient demographics [1]. However, existing adaptation methods face inherent clinical limitations. Specifically, Unsupervised Domain Adaptation (UDA) [7] and Test-Time Training (TTT) [20] rely on privacy restricted source data, whereas Fine-Tuning (FT) [21] and Continual Learning (CL) [6] demand exorbitantly costly target annotations. Furthermore, although Source-Free Domain Adaptation (SFDA) [14] bypasses these prerequisites, its offline optimization precludes instantaneous updates on continuous online data streams. In contrast, Test-Time Adaptation (TTA) [13] offers a superior "adapt-while-testing" paradigm. By directly ingesting unlabeled online streams during inference, TTA achieves simultaneous testing and updating without requiring source data or target labels, effectively resolving the bottlenecks of real-world clinical environments.

While TTA shows remarkable potential against distribution shifts, existing methods are often constrained by overly simplified optimization objectives. Early works like Tent [25] focus solely on minimizing predictive entropy, which under unknown shifts invariably induces dangerous overconfidence in erroneous predictions. To enhance robustness, methods like CoTTA [26] incorporate the Mean Teacher architecture [23,12] to enforce predictive consistency between teacher and student models. Yet, this uni-dimensional constraint often forces the model to maintain blind stability on already flawed representations. Recently, studies have attempted to jointly analyze both metrics. GraTa [5] aligns gradients from entropy and consistency to dynamically decay learning rates for unreliable samples, which essentially remains a passive avoidance strategy. Although EATA-C

[22] employs Min-Max regularization to actively penalize the entropy of high-risk samples, it strictly limits consistency to uni-directional minimization and relies on heuristics lacking systematic interpretability. Consequently, current explorations of entropy and consistency remain nascent and fragmented. To address this, we propose the Entropy-Consistency Quadrants (Fig. 1(a)). By integrating these disparate metrics into a cohesive analytical paradigm, this diagnostic framework provides interpretable and empirically grounded data-flow guidance for the safe clinical adaptation of FMs.

Based on the aforementioned analysis framework, we partition the streaming data into four quadrants, as illustrated in Fig. 1(a). The High Consistency & Low Entropy (HCLE) quadrant denotes the **Well Adapted** zone. Samples herein typically exhibit exceptionally high accuracy; thus, routing potentially correct predictions into this quadrant represents the ideal state of model adaptation. Conversely, the Low Consistency & High Entropy (LCHE) quadrant designates the **High Risk** zone because it exhibits the lowest empirical accuracy and the densest incorrect predictions in our zero-shot analysis, rather than merely reflecting an entropy-consistency correlation. Sequestering potentially incorrect samples into this region is paramount to prevent the model from learning noisy pseudo-labels. The remaining two quadrants (HCHE and LCLE) serve as transitional buffers. To validate the empirical relevance of this partitioning, we conducted a zero-shot preliminary study using the Ark+ [16] and DermLIP [28] FMs across three Out-of-Distribution (OOD) datasets. By mapping the prediction streams onto the proposed quadrants and partitioning samples by population medians of entropy and consistency for comparable quadrant statistics, we uncovered a robust accuracy hierarchy (Fig. 1(b)):

$$\text{Acc(HCLE)} > \text{Acc(HCHE and LCLE)} > \text{Acc(LCHE)} \tag{1}$$

This objective hierarchy reveals a crucial insight: the HCLE quadrant predominantly concentrates potentially correct samples, whereas the LCHE quadrant effectively isolates erroneous predictions. Driven by this pivotal finding and the pursuit of ideal adaptation flows, we formally propose the Entropy-Consistency Flow Rectification (ECFR) approach. Unlike standard paradigms such as Tent [25] and CoTTA [26] that rely on unidirectional blind minimization, ECFR introduces a quadrant-wise flow rectification. This establishes a bidirectional optimization explicitly driven by quadrant partitioning, thereby safely and precisely guiding FMs through clinical adaptation.

2 Methodology

We propose the Entropy-Consistency Flow Rectification (ECFR) approach to operationalize the quadrant-driven data flow strategy derived from the Entropy-Consistency Quadrants. As illustrated in Fig. 2, ECFR consists of an online state diagnosis module for dynamic quadrant partitioning, a quadrant-wise flow rectification mechanism for bidirectional optimization, and a dynamic memory bank for robust thresholding.

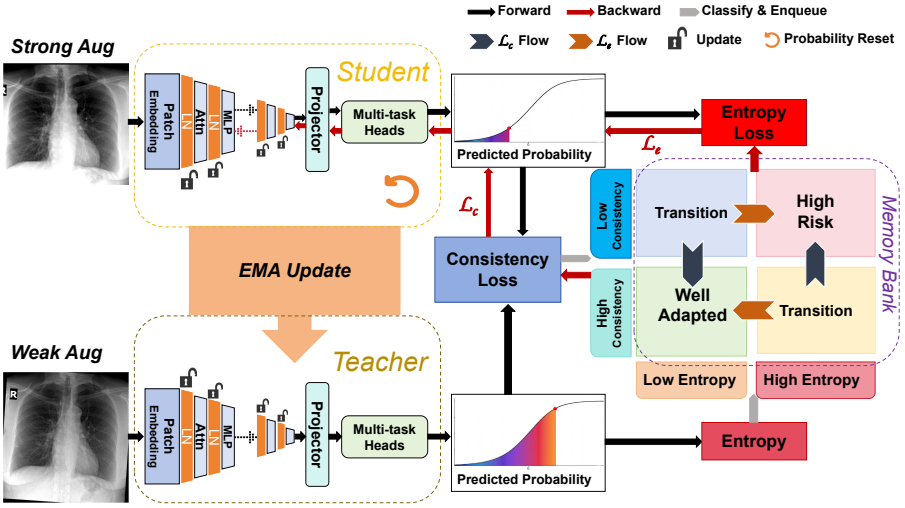


Fig. 2. Overview of the ECFR approach. A Mean Teacher architecture [23] generates weak (p_t) and strong (p_s) predictions. The quadrant division guides the bidirectional optimization to dynamically switch between minimization (consolidation) and maximization (sequestration), stabilized by a Dynamic Memory Bank.

2.1 Bidirectional Optimization Objectives

Based on the quadrants constructed in the Introduction, we formally define the optimization objectives and provide the heuristic rationale for the bidirectional strategy.

Metric Definition. We employ the Mean Teacher paradigm [23], maintaining a student model f_θ and an Exponential Moving Average (EMA) teacher $f_{\theta'}$. For an input $\boldsymbol{x}$, $\mathcal{T}_w$ and $\mathcal{T}_s$ denote weak and strong stochastic transformations of the same test image, and the corresponding predictions are $\boldsymbol{p}_t = f_{\theta'}(\mathcal{T}_w(\boldsymbol{x}))$ and $\boldsymbol{p}_s = f_\theta(\mathcal{T}_s(\boldsymbol{x}))$. To dynamically quantify the diagnostic state of each sample, we evaluate its predictive entropy (H) and consistency (D). Specifically, the entropy [19] of the student's prediction is calculated as $H(\boldsymbol{p}_s) = -\sum_c p_s(c) \log p_s(c)$. Concurrently, the consistency is quantified by the discrepancy between teacher and student predictions via the Jensen-Shannon (JS) divergence [15]: $D(\boldsymbol{p}_t, \boldsymbol{p}_s) = \frac{1}{2}[\mathcal{D}_{KL}(\boldsymbol{p}_t||\boldsymbol{m}) + \mathcal{D}_{KL}(\boldsymbol{p}_s||\boldsymbol{m})]$, where $\boldsymbol{m} = \frac{1}{2}(\boldsymbol{p}_t + \boldsymbol{p}_s)$ and $\mathcal{D}_{KL}$ is the Kullback-Leibler divergence [11]. Unlike standard KL divergence, JS divergence is symmetric and bounded in $[0, \ln 2]$, inherently preventing gradient explosion during subsequent maximization optimization.

Heuristic Motivation. We provide the heuristic motivation for our Min-Max strategy over conventional universal minimization. The rationale for bidirectional

entropy optimization is grounded in the Brier Score (BS) [4], defined as the mean squared error between predicted probabilities and true labels, i.e., $BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$. While minimizing entropy for potentially correct samples in the HCLE quadrant conventionally reduces the BS, this logic catastrophically fails for high-risk erroneous samples in the LCHE quadrant. For these LCHE samples, an overconfident incorrect prediction severely inflates the BS. Maximizing their entropy pushes the probability towards uniformity ($p \approx 0.5$), penalizing confident errors more safely than blind minimization.

Concurrently, the bidirectional optimization of consistency (JS divergence) is essential for balancing model plasticity and stability. Minimizing JS divergence for HCLE samples acts as standard knowledge distillation. However, applying this same minimization to LCHE samples exacerbates confirmation bias by blindly reinforcing flawed features [2]. Therefore, we actively maximize JS divergence for these high-risk samples to generate a repulsive gradient signal. Aligning with the Maximum Classifier Discrepancy (MCD) theory [18], this deliberate structural perturbation propels the model away from the current erroneous decision boundary, providing the momentum necessary to escape local optima.

2.2 Quadrant-Wise Flow Rectification

To operationalize the aforementioned heuristic motivations, we explicitly route the data flow based on dynamic quadrant partitioning. Accordingly, we formulate the total loss $\mathcal{L}_{total}$ as a dynamic linear combination of predictive entropy and JS divergence:

$$\mathcal{L}_{total} = \lambda_e \cdot H(\boldsymbol{p}_s) + \lambda_c \cdot D(\boldsymbol{p}_t, \boldsymbol{p}_s) \tag{2}$$

Crucially, λ_e and λ_c are not tunable loss-weight hyperparameters but gradient-direction switches: $+1$ minimizes the corresponding term, whereas -1 makes minimizing the signed objective equivalent to maximizing it. These switches are dictated by a sample's spatial position within the 2D diagnostic space, yielding the following conditional piecewise function:

$$\lambda_e = \begin{cases} 1 & \text{if } D < \tau_D \\ -1 & \text{if } D \geq \tau_D \end{cases}, \quad \lambda_c = \begin{cases} 1 & \text{if } H < \tau_H \\ -1 & \text{if } H \geq \tau_H \end{cases} \tag{3}$$

As illustrated in Fig. 2, this formulation steers samples toward the desired diagnostic quadrants. Specifically, samples falling into the HCLE zone inherently satisfy $H < \tau_H$ and $D < \tau_D$, thereby triggering dual minimization ($\lambda_e = 1, \lambda_c = 1$) to actively consolidate potentially correct predictions. Conversely, samples in the LCHE zone ($H \geq \tau_H, D \geq \tau_D$) strictly trigger dual maximization ($\lambda_e = -1, \lambda_c = -1$) to explicitly sequester high-risk erroneous features and halt confirmation bias. For transitional states (HCHE and LCLE), the mixed coefficients induce an adversarial optimization that naturally propels samples toward these appropriate extremes, yielding interpretable flow rectification for safe clinical adaptation.

2.3 Adaptive Protocols for Streaming Data

To combat distribution drift, we maintain a First-In-First-Out (FIFO) queue as a memory bank $\mathcal{B}$ (capacity N) to archive recent entropy and JS divergence values. Dynamic thresholds are calibrated in real time via history quantiles: $\tau_H = Q(\mathcal{B}_H, q)$ is the q-quantile of recent entropy values, and $\tau_D = Q(\mathcal{B}_D, q)$ is the q-quantile of recent JS-divergence values. The hyperparameter q acts as a tolerance prior governing diagnostic stringency, formulating our generic *ECFR-q* strategy (detailed in Sec. 3.1).

Following recent findings that optimizing a minimal parameter fraction is sufficient for adapting high-performing FMs [3,27], we exclusively update the LayerNorm affine parameters. Additionally, Stochastic Restoration [26] acts as a source-anchoring regularizer by resetting a small fraction of trainable student parameters to their pre-trained values during online adaptation.

3 Experiments

3.1 Experimental Setup

Foundation Models and Target OOD Datasets. To evaluate ECFR across distinct modalities, we employ the official pretrained weights of two FMs alongside three strictly unseen OOD datasets exhibiting realistic clinical shifts. For chest radiography, we utilize **Ark+** [16] (pretrained on over 700K adult X-rays) and adapt it to **Mendeley-V2** [9] (5,856 images), which introduces a severe demographic shift from adult to pediatric patients. For dermatology, we adopt **DermLIP** [28], a vision-language FM pretrained on the million-scale Derm1M dataset. Its adaptation is evaluated on two distinct datasets: **Daffodil** [10] (9,548 images, 5 non-melanoma/rare skin diseases) and **HAM10000** [24] (10,015 images, 7 pigmented lesions).

Implementation Details. We simulate real-time clinical deployment via a strict online streaming protocol (batch size = 1, no data replay). LayerNorm affine parameters are exclusively optimized via SGD, with learning rates determined by a quasi-logarithmic grid search ($\{1, 3.3, 6.6\} \times \{10^{-5}, 10^{-4}\} \cup \{10^{-3}\}$) for fair comparison. A FIFO memory bank ($N = 128$) computes online quantile thresholds, while Stochastic Restoration ($p = 0.01$) [26] limits drift but may suppress adaptation if set too large. We evaluate ECFR-0.5 ($q = 0.5$) as a no-prior setting and ECFR-Acc as a prior-informed variant, where q uses an overall target-domain reliability estimate from deployment logs, institutional validation, or a small audit cohort; in our public benchmarks, zero-shot accuracy is used only as this scalar proxy, with no per-sample labels during online adaptation. Efficacy and safety are assessed via Brier Score (BS) [4], Expected Calibration Error (ECE, 20 bins) [8], and Accuracy (Acc).

Table 1. Comparison of adaptation performance across three adaptation scenarios. **BS**: Brier Score (↓); **ECE**: Expected Calibration Error (↓); **Acc**: Accuracy (↑). Best results are **bold**, second-best underlined.

Method	Ark+ → Mendeley-V2			DermLIP → Daffodil			DermLIP → HAM		
	BS ↓	ECE ↓	Acc ↑	BS ↓	ECE ↓	Acc ↑	BS ↓	ECE ↓	Acc ↑
Zero-Shot	0.4116	0.5270	0.3185	0.3829	0.1840	0.7921	0.6882	0.3276	0.6269
Tent [25]	0.4954	0.5873	0.2814	0.3851	0.1847	0.7919	0.5427	0.2584	0.7070
CoTTA [26]	0.4209	0.5343	0.3004	0.3413	0.1647	0.8151	0.5188	0.2510	0.7206
SAR [17]	0.4577	0.5617	0.2865	0.3278	0.1548	0.8199	0.5923	0.2839	0.6803
GraTa [5]	0.4476	0.5542	0.2876	0.3870	0.1849	0.7926	0.5218	0.2535	0.7230
EATA-C [22]	0.4092	0.5259	0.3128	0.3339	0.1654	0.8299	0.6859	0.3264	0.6289
ECFR-0.5	0.3893	0.4996	0.3227	0.3184	0.1519	0.8335	0.5208	0.2508	0.7179
ECFR-Acc	**0.3173**	**0.4272**	**0.4173**	**0.2799**	**0.1338**	**0.8468**	**0.5107**	**0.2439**	**0.7232**

3.2 Main Results

Comparative Experiments and Metric Evaluation. Table 1 benchmarks ECFR against a Zero-Shot baseline, which serves as the no-adaptation reference, and five TTA methods: Tent [25], CoTTA [26], SAR [17], GraTa [5][5], and EATA-C [22]. Under severe shifts (e.g., Ark+ → Mendeley-V2), existing paradigms suffer from error accumulation, where blind metric minimization degrades performance below the Zero-Shot baseline. Conversely, both ECFR variants demonstrate consistent robustness. The prior-guided ECFR-Acc achieves the best results across all metrics and datasets, while the target-agnostic ECFR-0.5 stably secures most second-best outcomes. This verifies that bidirectional optimization effectively prevents model collapse during adaptation. Fully supervised target-domain training is not included as an upper-bound comparator because it uses target labels and is not directly comparable to unsupervised online TTA.

Data Flow Visualization. To elucidate the underlying data flow dynamics, Fig. 3 maps the predictive trajectories of Tent [25], CoTTA [26], and ECFR onto our proposed Entropy-Consistency Quadrants. Within this framework, both baselines expose the critical flaw of unidirectional blind optimization: they indiscriminately force erroneous samples into low-metric quadrants, severely exacerbating confirmation bias. While existing TTA methods naturally exhibit varied flow trajectories, ECFR pioneers a quadrant-wise flow rectification that explicitly sequesters high-risk erroneous predictions into the LCHE zone while consolidating potentially correct samples within the HCLE quadrant. Ultimately, these visualizations empirically instantiate the ideal adaptation flow postulated earlier, aligning the proposed heuristic with observable dynamics and corroborating ECFR's superiority in halting error accumulation.

[5] Originally tailored for segmentation, we adapted GraTa for classification tasks.

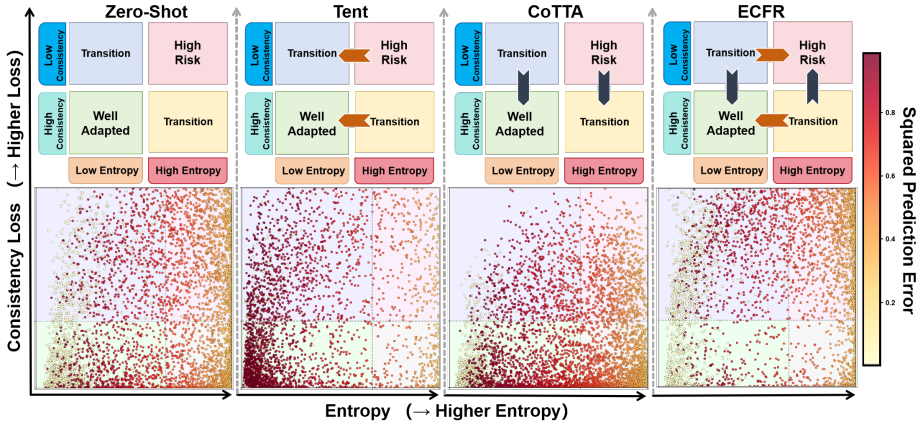


Fig. 3. Optimization dynamics via Entropy-Consistency data flow. Scatter plots (X: Entropy, Y: Consistency) visualize sample shifts (Color: Squared Prediction Error; Red: High Error). We employ ECFR-0.5 for visualization here, as the quadrants are delineated by the population medians of entropy and JS divergence derived from the zero-shot baseline on the Mendeley-V2 dataset.

3.3 Ablation Study

Table 2 demonstrates that ablating either the entropy-guided (w/o Ent) or consistency-guided (w/o Cons) low routing consistently degrades adaptation performance. The optimal results achieved by the full ECFR strategy across all datasets confirm the essential synergistic nature of integrating both metrics for precise quadrant-wise flow rectification.

Table 2. Ablation study on the flow rectification components using ECFR-Acc. $\mathcal{L}_\mathbf{e}$: Entropy-guided flow; $\mathcal{L}_\mathbf{c}$: Consistency-guided flow. Format as Table 1.

Modules		**Ark+ → Mendeley-V2**			**DermLIP → Daffodil**			**DermLIP → HAM**		
$\mathcal{L}_\mathrm{e}$	$\mathcal{L}_\mathrm{c}$	BS ↓	ECE ↓	Acc ↑	BS ↓	ECE ↓	Acc ↑	BS ↓	ECE ↓	Acc ↑
✗	✗	0.4116	0.5270	0.3185	0.3829	0.1840	0.7921	0.6882	0.3276	0.6269
✗	✓	0.4197	0.5333	0.3023	<u>0.3485</u>	<u>0.1662</u>	<u>0.8101</u>	<u>0.5151</u>	<u>0.2488</u>	<u>0.7223</u>
✓	✗	<u>0.3183</u>	<u>0.4300</u>	<u>0.4112</u>	0.3577	0.1705	0.8057	0.5529	0.2649	0.7033
✓	✓	**0.3173**	**0.4272**	**0.4173**	**0.2799**	**0.1338**	**0.8468**	**0.5107**	**0.2439**	**0.7232**

4 Conclusion

We present ECFR, a test-time adaptation method that safely deploys medical FMs under continuous distribution shifts. Guided by Entropy-Consistency Quadrants, ECFR introduces a bidirectional optimization with dynamic threshold calibration to consolidate reliable predictions and isolate high-risk samples. This heuristic and interpretable flow rectification mitigates confirmation bias and prevents model collapse, achieving superior safety across three challenging clinical OOD benchmarks. Future work will extend validation to more model architectures and disease datasets, and further develop the methodology.

Acknowledgments. This work is supported by National Key R&D Program of China (2022ZD0212400),Natural Science Foundation of China (62373243), Shanghai Municipal Health Commission Smart Healthcare Project (2025ZHYL021), Natural Science Foundation of Shanghai (25ZR1402276), Open Project of the Shanghai Key Laboratory of Flexible Medical Robotics (Grant No. SKLFMR-0211), and Shanghai Tongren Hospital Medical-Engineering Collaboration Project (lhyjzx2024-xm03).

Disclosure of Interests. The authors declare no competing interests.

References

1. Ahmad, I.S., Suleiman, R.B., Yu, T., Lin, R., Zhao, C., Liao, J., Wu, B., Xie, Y., Liang, X., Wang, H., Hu, Z.: Foundation models for x-ray interpretation: a narrative review of current techniques and future perspectives in diagnostic imaging. Quantitative Imaging in Medicine and Surgery **16**(4), 321 (2026). https://doi.org/10.21037/qims-2025-1-2782
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020). https://doi.org/10.1109/IJCNN48605.2020.9207304
3. Basu, S., Massiceti, D., Hu, S.X., Feizi, S.: Strong baselines for parameter-efficient few-shot fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 11024–11031 (2024). https://doi.org/10.1609/aaai.v38i10.28978
4. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1–3 (1950). https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
5. Chen, Z., Ye, Y., Pan, Y., Xia, Y.: Gradient alignment improves test-time adaptation for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 39, pp. 2429–2437 (2025). https://doi.org/10.1609/aaai.v39i3.32244
6. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3366–3385 (2022). https://doi.org/10.1109/TPAMI.2021.3057446

7. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2022). https://doi.org/10.1109/TBME.2021.3117407
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017), https://proceedings.mlr.press/v70/guo17a.html
9. Kermany, D.: Labeled optical coherence tomography (oct) and chest x-ray images for classification (2018). https://doi.org/10.17632/rscbjbr9sj.2
10. Khushbu, S.A.: Skin disease classification dataset (2024). https://doi.org/10.17632/3hckgznc67.1
11. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694
12. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (ICLR) (2017), https://openreview.net/forum?id=BJ6oOfqge
13. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025). https://doi.org/10.1007/s11263-024-02181-w
14. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 6028–6039 (2020), https://proceedings.mlr.press/v119/liang20a.html
15. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
16. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open ai foundation model applied to chest radiography. Nature **643**, 488–498 (2025). https://doi.org/10.1038/s41586-025-09079-8
17. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=g2YraF75Tj
18. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3723–3732 (2018). https://doi.org/10.1109/CVPR.2018.00392
19. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
20. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A.A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 9229–9248 (2020), https://proceedings.mlr.press/v119/sun20b.html
21. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Transactions on Medical Imaging **35**(5), 1299–1312 (2016). https://doi.org/10.1109/TMI.2016.2535302

22. Tan, M., Chen, G., Wu, J., Zhang, Y., Chen, Y., Zhao, P., Niu, S.: Uncertainty-calibrated test-time model adaptation without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(8), 6274–6289 (2025). https://doi.org/10.1109/TPAMI.2025.3560696
23. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
24. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**, 180161 (2018). https://doi.org/10.1038/sdata.2018.161
25. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (ICLR) (2021)
26. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7201–7211 (2022). https://doi.org/10.1109/CVPR52688.2022.00706
27. Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., Huang, Z.: In search of lost online test-time adaptation: a survey. International Journal of Computer Vision **133**(3), 1106–1139 (2025). https://doi.org/10.1007/s11263-024-02213-5
28. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H., Tschandl, P., Kittler, H., Ge, Z.: Derm1m: a million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12681–12690 (2025). https://doi.org/10.1109/ICCV51701.2025.01178